Институт за математику и информатику
Природно-математички факултет
Универзитет у Крагујевцу

Мастер рад

Примена алгоритма моделовања тема из текстуалних садржаја за проналажења одговора на постављено питање

Студент:
Јелица Васиљевић

Професор:
др Милош Ивановић

Август 2015.

Глава 1

Математичка позадина

У претходним поглављима, рад се углавном бавио питањима *шта је TM алгоритам и чему служи*, без улажења у то ***како*** он уствари ради.

Опис рада TM алгоритама - конкретно LDA имплементације, биће изложен у неколико целина. Најпре ће се објаснити (увести) неки појмови вероватноће који су битни за разумевање сусштине рада алгоритма, а затим ће бити изнешена математичка позадина самог алгоритма.

1.1 Теорија вероватноће

Теорија вероватноће је математичка дисциплина која се бави изучавањем случајних појава тј. појава чији исходи нису увек строго дефинисани.

Први проблеми који се могу сматрати проблемима вероватноће потичу још из 12. века и везани су за проучавање исхода разних игара на срећу. Развој *теорије вероватноће* почиње средином 17. века и везан је за имена Блеза Паскала,Пјера де Ферма и Кристијана Хајгенса. Наиме, између Паскала и Ферма је 1654. године започела интересантна преписка о низи проблема међу којима је био и проблем везан за поделу улога приликом прекида једне коцкарске игре. Проблем је био постављен на следећи начин : Два играча А и Б се договоре да читав улог припадне ономе ко први добије три игре. Када је играч А добио 2 игре а играч Б једну игру, играчи су споразумно одлучили да прекину игру. Поставља се питање како сада да поделе улог. Паскал је предложио поделу у односу 3:1 у корист играча А. Овај пример често се узима као почетак настанка теорије вероватноће.

Неке од појава које се догађају у реалном свету лако се могу предвидети и објаснити услод познавања законитости њиховог настанка. У такве појаве спадају нпр. помрачење Сунца и Месеца, плима и осека, гравитација итд. Међутим, постоје појаве чије узроке тренутно није могуће одредити па се не могу у потпуности објаснити и одредити. Неке од таквих појава су нпр. добитак на лутрији или метереолошке појаве. Прилоком бацања металног, хомогеног новчића, никада није сигрно да ли ће пасти писмо или глава. Међутим, уколико бацамамо новчић много пута, може се уочити да је отприлике исти број пута пало писмо као и глава (такве експерименте су радили Буфон и ПирсонДакле, законитост код оваквих догађаја може се уочити тек након великог броја понављања појаве.

1.1.1 Основни појмови

Основни полазни појам у теорији вероватноће је непразан скуп Ω који представља скуп свих могућих исхода једног експеримента. Овај скуп се често назива **простор елементарних догађаја** и може бити коначан, пребројив или непребројив. **Случајни догађаја** или само **догађај** представља било који подскуп скупа Ω. Најчешће се случајни догађаји означавају великим, штампаним, латиничним словима. За догађај **A** каже се да се **реализовао** ако

се реализовао неки исход ω који припада скупу А. Догађај који је садржи све могуће елементарне исходе експеримента назива се **сигуран догађај** а догађај који не садржи ни један елементарни исход назова се **немогућ догађај**.

Пример: Нека је дата хомогена коцка чије су стране означене бројевима од 1 до 6. Елементарни догађаји су појављивање одређеног броја при бацању коцкице. Према томе, скуп свих могућих исхода екперимента бацања коцкице је $\Omega = \{1, 2, 3, 4, 5, 6\}$. Догађај А = "пао је паран број"одређује скуп А= $\{2, 4, 6\}$

Производ два догађаја и , у ознаци AB је догађај који се реализује ако и само се ако реализују оба догађаја. Дакле, производ догађаја је пресек скупова А и В. Уколико су А и В дисјунктни скупови (пресек је празан скуп) за такве догађаје кажемо да су **несагласни** или да се **искључују**.

Збир два додгађаја А и В, у ознаци $A \cup B$ представља догађај који се реализује ако се реализује бар један од догађаја А и В.

Разликом догађаја А и В, у ознаци $A - B$ назива се догађај који се реализује ако и само ако се реализује догађај А а не реализује догађај В.

Потпун систем догађаја : За догађаје $A_1, A_2, ..A_n$ се каже да образују *потпун систем догађаја* уколико важи : $\bigcup_i A_i = \Omega$. Дакле, при реализацији неког експеримента бар један од догађаја $A_1, A_2, ..A_n$ ће се реализовати. Посебно су интересантни потпуни системи несагласних догађаја као што се може видети код формуле тоталне вероватноће.

Дефиниција 1 (Класична дефиниција вероватноће) *: Нека је $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ скуп свих могућих једнаковероватних елементарних догађа који су међусобно несагласни и нека је $A = \{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_m}\}$ догађај који се састоји од m елементарних једнаковероватних догађаја. Вероватноћа наступања догађаја A је :*

$$P(A) = \frac{m}{n} \tag{1.1}$$

Претходна дефиниција може се неформално изразити и овако : вероватноћа догађаја А једнака је количнику броја **повољних исхода** експеримена (исходи када се реализује догађај А) и укупног броја свих могућих исхода експеримента.

Класична дефиниција вероватноће је применљива само онде где су елементарни догађаји једнаковероватни. Међутим, тај услов је у пракси јако тешко испунити. Чак и у случајевима када је то наизглед очигледно, као што је бацање коцкице, једнаковероватност не може бити гарантована. Разлози за то могу бити технологија израде коцкице која не мора бити савршено прецизна, немогућност обезбеђивања идеалних и непромењљивих услова током извођења експеримента итд. Због тога је једини начин којим је могуће заиста утврдити вероватноћу догађаја А *статистички приступ* заснован на великом броју експеримената.

Дефиниција 2 (Статистичка дефиниција вероватноће) *: Нека се у n понављања експеримента изведених под приближно истим условима догађај A реализовао m_n пута. Вероватноћа догађаја A је*

$$P(A) = \lim_{n \to \infty} \frac{m_n}{n} \tag{1.2}$$

1.1.2 Условна вероватноћа

Вероватноћа догађаја чија реализација **не зависи** од наступања било ког другог догађаја назива се **безусловна вероватноћа**. Ако је реализација догађаја А условљена реализацијом неког догађаја В при чему В није немогућ догађај ($P(B) \neq 0$), тада се вероватноћа догађаја под условом да се десио догађај В назива **условном вероватноћом** и означава се са $P(A \mid B)$. Дакле, $P(A \mid B)$ је вероватноћа догађаја А под условима који сигурно доводе до реализације догађаја В.

Нека се изводи експеримент у коме постоји n једнаковероватних елементарних догађаја и нека је са n_A, n_B, n_{AB} означен број елеменгтарних догађаја који доводе до реализације догађаја A, B, AB редом.

Према класничној дефиницији вероватноће, вероватноћа реализације догађаја A и AB је :

$$P(B) = \frac{n_B}{n}, P(AB) = \frac{n_{AB}}{n} \tag{1.3}$$

Ако је реализација догађаја A условљена реализацијом догађаја B, то је број повољних исхода догађаја A n_{AB} (број елементарних догађаја који имају осбине и скупа A и скупа B). Пошто се догађај A реализује само ако се реализовао догађај B, број свих могућих исхода је n_B (број свих могућих елементарних догађаја када наступа догађај B). Дакле, условна вероватноћа догађаја A, под условом да се десио догађај B је :

$$P(A \mid B) = \frac{n_{AB}}{n_B} = \frac{\frac{n_{AB}}{n}}{\frac{n_B}{n}} = \frac{P(AB)}{P(B)} \quad , P(B) \neq 0 \tag{1.4}$$

У случају да је догађај B условљен догађајом A, аналогно се изводи да је

$$P(B \mid A) = \frac{P(AB)}{P(A)} \quad , P(A) \neq 0 \tag{1.5}$$

Из релација (3.4) и (3.5) следи

$$P(AB) = P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \tag{1.6}$$

Релација (3.6) назива се још и **теорема о производу вероватноћа**

Теорема 1 (Формула тоталне вероватноће) *: Ако су $H_1, H_2, .., H_n$ међусобно несагласни догађаји, $P(H_i) > 0 (i = 1, .., n)$ при чему важи $H_1 + H_2 + .. + H_n = \Omega$ тада је :*

$$P(A) = \sum_{i=1}^{n} P(H_i) P(A \mid H_i) \text{ за сваки догађај} A \subseteq \Omega \tag{1.7}$$

Напомена : Догађаји $H_1, H_2, .., H_n$ чине потпун систем несагласних догађаја.

Доказ 1 *Обзором да су догађаји подскупови скупа свих елементарних догађаја очигледно је да важи*

$$A = A\Omega = A \sum_{i=1}^{n} H_i = \sum_{i=1}^{n} AH_i. \tag{1.8}$$

На основу релације (3.6) следи :

$$P(A) = P(\sum_{i=1}^{n} AH_i) = \sum_{i=1}^{n} P(AH_i) = \sum_{i=1}^{n} P(H_i) P(A \mid H_i) \tag{1.9}$$

Вероватноће $P(H_i)$ су обично познате унапред и називају се **априорним вероватноћама** а сами догађаји **хипотезама**.

Теорема 2 (Бајесова формула [1]) *: Ако су $H_1, H_2, .., H_n$ међусобно несагласни догађаји, $P(H_i) > 0 (i = 1, .., n)$ при чему важи $H_1 + H_2 + .. + H_n = \Omega$ тада је :*

$$P(H_i \mid A) = \frac{P(H_i) P(A \mid H_i)}{\sum_{j=1}^{n} P(H_j) P(A \mid H_j)} \quad (i = 1...n) \quad \text{за сваки догађај} A \subseteq \Omega \tag{1.10}$$

Доказ 2 *Из релације (3.6) следи :*

$$P(H_iA) = P(H_i)P(A \mid H_i) = P(A)P(H_i \mid A) \quad (i = 1...n) \tag{1.11}$$

Условна вероватноћа догађаја H_i под условом да се десио догађај A је:

$$P(H_i \mid A) = \frac{P(H_iA)}{P(A)} = \frac{P(H_i)P(A \mid H_i)}{P(A)}$$

Примењујући формулу потпуне вероватноће за $P(A)$ добија се

$$P(H_i \mid A) = \frac{P(H_i)P(A \mid H_i)}{\sum_{j=1}^{n} P(H_j)P(A \mid H_j)}$$

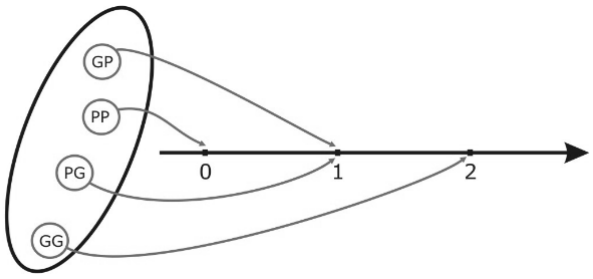
што представља Бајесову формулу.

1.1.3 Случајне променљиве

Ако се сваком елементарном догађају придружи један реалан број, онда се извођење експеримента може посматрати као избор вредности једне променљиве. Променљива величина која те бројене вредности узима са одређеним вероватноћама назива се *случајна променљива*. Дакле, уместо вербалне карактеризације догађаја (описа речима шта догађај представља) много је једноставније за рад догађаје окарактерисати бројним вредностима тј. неким реалним бројевима.

Пример 1 : У експериманту бацања новчића могућа су два елементарна исхода : грб или писмо. Нека је догађај који се посматра "пало је писмо". Појава писма се може означити бројем 1 а појава грба бројем 0. Сада се овај експеримент може замислити као избор 0 или 1 са вероватноћом $\frac{1}{2}$

Пример 2 : Новчић се баца два пута. Нека је са P означена појава писма а са G појава грба. Скуп свих елементарних исхода експеримента је $\Omega = \{PP, PG, GP, PP\}$. Нека је догађај који се посматра "број палих писама". Сваком исходу се може доделити један реалан број и то $PP- > 2, GP- > 1, PG- > 1, GG- > 0$. Ово додељивање вредности се карактерише случајном променљивом. Случајна променљива сваку од ових вредности узима са различитом вероватнићом.



Слика 1.1: Графички пример случајне променљиве

Дефиниција 3 *Функција X која сваком случајном догађају $\omega \in \Omega$ додељује неки реалан број $X(\omega)$ назива се* ***случајна променљива*** *где је $X : \Omega \longrightarrow R$*

Дакле, случајна променљива је пресликавање скупа Ω у скуп **реалних** бројева за разлику од вероватноће која је пресликавање скупа Ω у скуп $[0, 1]$

Важно је уочити да случајна променљива **нема конкретну вредност** већ се само говори о вероватноћама да узме неки конкретну вредност.

Разликују се два основна типа случајних променљивих - **дискретне** и **непрекидне**. Подела се врши у зависности од тога да ли случајна променљива узима вредности у пребројивом или непребројивом скупу вредности.

1.1.3.1 Дискретне случајне промењиве

За случајну промењиву се каже да је дискретног типа ако узима коначан број изолованих вредности или пребројиво много вредности

Дефиниција 4 *Нека случајна промењива X може да узме вредности $x_1, x_2, ..., x_n$ са вероватноћама $p_1, p_2, ..., p_n$ при чему важи да је $p_1 + p_2 + ... + p_n = 1$. Скуп парова $(x_i, p_i = P\{X = x_i\}), i = 1, 2, ..., n$ или написано :*

$$\begin{pmatrix} x_1 & x_2 & \cdots \\ p(x_1) & p(x_2) & \cdots \end{pmatrix}$$

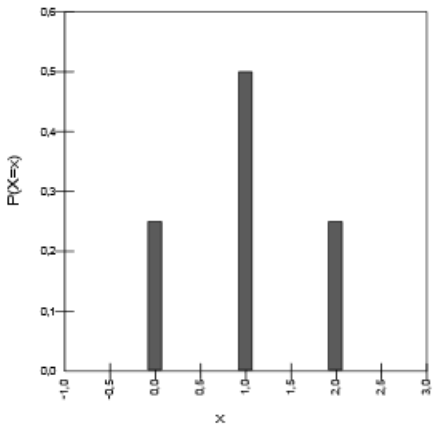
чине закон расподеле или распоред вероватноћа случајне промењиве X.

Закон расподеле случајне промњиве може да се посматра као **правило** по коме се свакој вредности случајне промењиве придружује одговарајућа вероватноћа. Дакле, при реализацији експеримента сигурно ће се десити догађај којем је придужена нека вредност случајне промењиве. Због тога је сума свих вероватноћа у расподели случајне промењиве 1. Међутим, нису све вредности подједнако вероватне па се свакој вредноти придружује вероватноћа са којом се очекује. Претходна дефиниција може се интерпретирати и на следећи начин : извесна маса једнака јединици је распоређена на такав начин да се у тачкама $x_1, x_2, ..., x_n$ налазе одговарајући делови масе $p_1, p_2, ..., p_n$. Услед оваквог тумачења, закон расподеле вероватноћа се често назива и **функција масе вероватноћа**

У примеру 2, случајна промењива може да узме три вредности, тј. писмо се може појавити 0,1 или 2 пута у два бацања. Ни један други исход није могућ - нпр. у два бацања писмо не може да се појави 3 пута или -1 пут. Међутим, вероватноћа да се писмо неће појавити ни једном (или да се појави два пута) је $\frac{1}{4}$ - вероватноћа да падне глава у првом бацању је $\frac{1}{2}$ и вероватноћа да падне глава у другом бацању је $\frac{1}{2}$, дакле, вероватноћа да оба пута падне глава је је $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, вероватноћа да се писмо појави тачно једном је $\frac{1}{2}$ - писмо пада тачно једном у случају PG или GP. Вероватноћа за оба ова догађаја је $\frac{1}{4}$. Дакле, вероватноћа да се десио бар један од ових догађаја је $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Према томе, расподела случајне промењиве "број појављивања писма у два бацања "је :

$$\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Закон расподеле дискретне случајне промењиве може се представити графички, као на следећој слици :



Слика 1.2: Графички пример случајне промењиве

На апсциси се налазе могуће вредности случајних променљивих док се на ординати налазе вероватноће са којом случајна променљива узима дату вредности. Са претходне слике јасно се уочава дискретност посматране случајне променљиве - вероватноћа да случајна промељива узме вредност између неке две целобројне вредности је 0.

Функција расподеле дискретне случајне променљиве:

Распоред или закон расподеле случајне променљиве дискретног типа може се представити као листа свих могућих вредности случајне променљиве и одговарајућих вероватноћа. Међутим, поставља се питање како представти случајну променљиву која може узимати јако пуно вредности тј. бесконачно много вредности. У овом случају би требало формирати листу од бесконачно много чланова, што је практино неизводљиво. (Пример једне такве случајне променљиве би био - број бацања коцкице док се не добију две узастопне шестице. Случајна променљива може узети вредности 2,3,4,... са различитим вероватноћама, при чему не постоји горња граница броја бацања при којој се сигурно добијају две узастопне шестице). Због описаног проблема, потребно је пронаћи другачији начин представљања случајне променљиве и одговарајућих вероватноћа. То се постиже **функцијом расподеле** која се може дефинисати за сваку случајну променљиву.

Дефиниција 5 ***Функција расподеле*** *(још се назива и* ***кумулативна функција расподеле****) дискретне случајне променљиве претставља вероватноћу да случајна променљива X узме вредност која је мања или једнака неком реалном броју x при чему је дефинисана за свако реално x.*

$$F(x) = P(X < x) \quad \forall x \in R$$

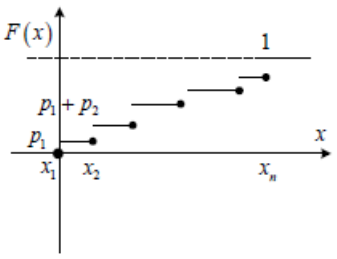
Дакле, кумулативна функција расподеле има облик

$$F(x) = \begin{cases} 0, & x \leq x_1 \\ p_1, & x_1 < x \leq x_2 \\ p_1 + p_2, & x_2 < x \leq x_3 \\ .... & ... \\ 1 & x > x_n \end{cases}$$

и може се изразити као :

$$F(x) = \sum_{k, x_k < x} P(X = x_k)$$

а графички приказ је дат на следећој слици :



Слика 1.3: График кумулативне функције расподеле случајне променљиве дискретног типа

Две најважније дискретне расподеле су **Биномна** и **Пуасонова** расподела.

1.1.3.2 Непрекидне случајне променљиве

Случајна променљива је (апсолутно) непрекидног типа ако може да узме **било коју** вредност из неког интервала. Број вредности које може да узме случајна променљива непрекидног типа је **бесконачан**. Неки од примера су : висина и тежина људи, дужина трајања батерије итд. На пример, нека је X случајна променљива која представља дужину рада сијалице. Ова случајна променљива може узети било коју вредност на интервалу од 1 до нпр. 1000 сати. Како у интервалу $[0, 1000]$ има бесконачно много реалних бројева, не постоји начин да се дефинише вероватноћа за сваку појединачну вредност, као што је био случај код дискретних променљивих. Такође, интуитивно је јасно да је вероватноћа да ће сијалица прегорети у тачно одређеном тренутку једнака 0 док је вероватноћа да ће прегорети у неком временском интервалу различита од нуле.

Дефиниција 6 *Случајна променљива X је апсолутно непрекидног типа ако постоји* ***ненегативна*** *функција* $f : \mathbb{R} - > \mathbb{R}$ *таква да за било који интервал* $[a, b] \subset (-\infty, \infty)$ *важи :*

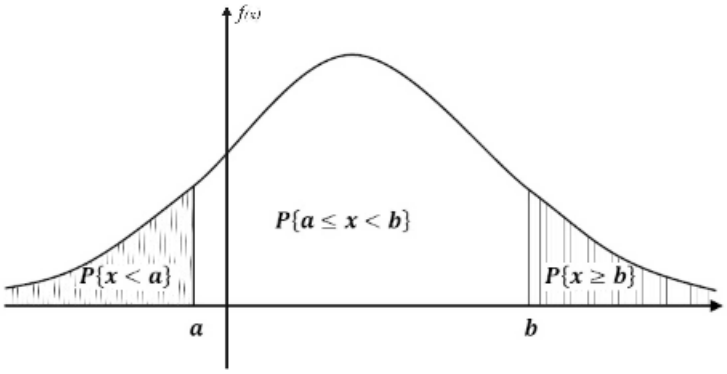
$$P\{a \leq X < b\} = \int_a^b f(x)dx \tag{1.12}$$

Функција $f(x)$ мора да задовољи услов :

$$P\{-\infty \leq X < \infty\} = P\{\Omega\} = \int_{-\infty}^{\infty} f(x)dx = 1$$

Функција $f(x)$ се назива **густина расподеле вероватноће** случајне променљиве X. Случјане променљиве **дискретног типа** немају густину расподеле баш као што ни случајне променљиве непрекидног типа немају закон расподеле вероватноћа.

Из релације (3.12) следи да је вероватноћа да случајна променљива узме вредност из интервала $[a, b]$ једнака **површини** испод графика функције $f(x)$ на интервалу $[a, b]$.



Слика 1.4: Функција густине

Функција расподеле непрекидне случајне променљиве:

Дефиниција 7 ***Функција расподеле (кумулативна функција расподеле)*** *непрекидне случајне променљиве се може представити као :*

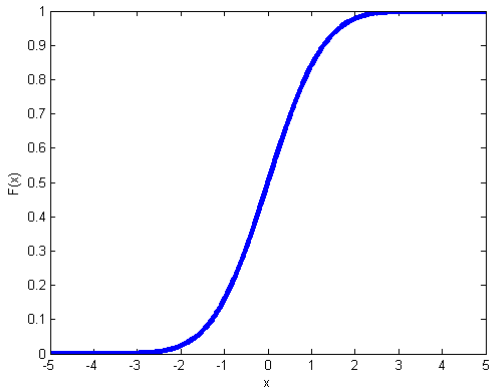
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt) \quad x \in (-\infty, \infty)$$

где је $f(x)$ *функција густине.*

Дефиниција кумулативне суме преко интеграла је јаснија ако се има на уму интервал из ког случајна променљива може да узме вредности. Код случајних променљивих дискретног типа, тај скуп је био пребројив па се кумулативна функција расподеле дефинисала преко суме. Случајне променљиве непрекидног типа могу узти бесконачно много вредности па се сума код дискретних случајних променљивих (када број тачака тежи у бесконачност) замењује интегралом.

Напомена : Ако случајна променљива X не узима све вредности из интервала $(-\infty, \infty)$ усваја се да је $f(x) = 0$ за све вредносто x из интервала у којима X не узима вредности.

График кумулативне функције расподеле непрекидне случајне променљиве X је сада представљен глатком кривом линијом (за разлику од случајне променљиве дискретног типа где је график био "степенаст").



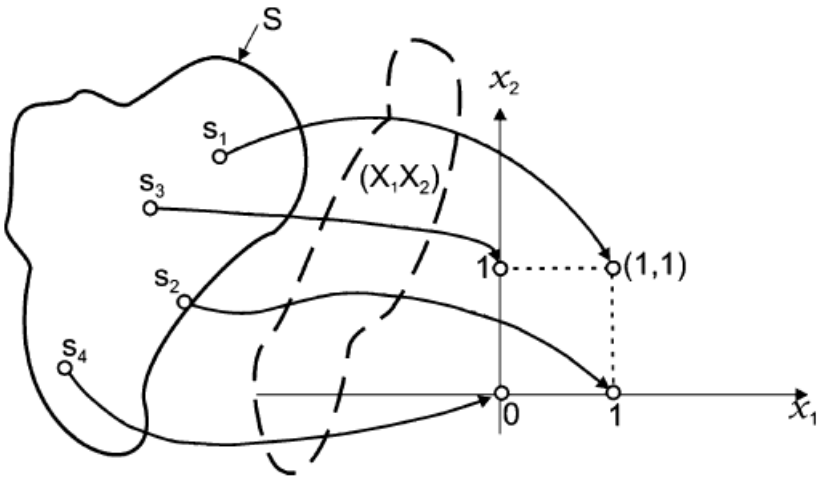
Слика 1.5: Кумулативна функција расподеле
за случајне променљиве непрекидног типа

1.1.3.3 Вишедимензионалне случајне прменљиве

Случајна променљива представља пресликавање скупа догађаја у реалне бројеве. Дакле, излази експеримента се мапирају у једнодимензионалан простор реалних бројева. Међутим, постоје случајеви када је потребно излазе експеримента мапирати у вишедимензионалне реалне просторе. На пример, при истовременом бацању два новчића могућа су 4 исхода :

1. s_1: први новчић писмо - други новчић писмо
2. s_2: први новчић писмо - други новчић глава
3. s_3: први новчић глава - други новчић писмо
4. s_4: први новчић глава - други новчић глава

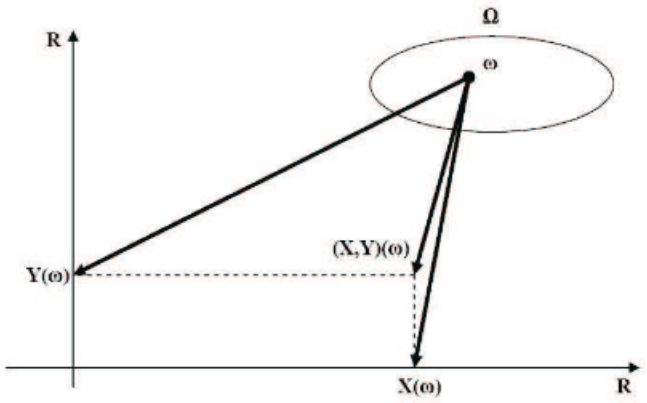
Нека је са X_1 означена случајна променљива која узима вредност 1 ако се на првом новчићу појавила глава, односно 0 ако се појавило писмо и аналогно X_2 која на исти начин означава појаву главе на другом новчићу. Исход експеримента се сада може описати дводимензионалном променљивом (X_1, X_2). Графички приказ ове дводимензионалне промељиве дат је на следећој слици :



Слика 1.6: Експеримент : бацање два новчића.
Скуп S представља скуп свих елементарних исхода (Ω)

Дефиниција 8 *Ако су* $X : \Omega -> \mathbb{R}, Y : \Omega -> \mathbb{R}$ *случајне променљиве, тада се* ***уређени пар*** (X, Y) *натива* ***дводимензионална случајна променљива****. Уређеним паром* (X, Y) *се сваком исходу* $\omega \in \Omega$ *придружује уређени пар бројева* $(X(\omega), Y(\omega)) = (x, y) \in \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

На следећој слици графички је представљена дводимензионална случајна променљива.



Слика 1.7: Дводимензионална случајна променљива

Овако уведен појам дводимензионалне случајне променљиве се може проширити и на више димензија и тада настају n-димензионалне случајне променљиве. Закључци изведени за дводимензионалне се такође односе и на вишедимензионалне случајне приомењиве.

Кумулативна функција расподеле (још се назива и *заједничка расподела енг. joint distribution* дводимензионалне случајне променљиве, у ознаци $F_{XY} : \mathbb{R}^2 -> [0, 1]$ дефинише се као вероватноћа реализације догађаја $\{X \leq x, Y \leq y\}$ односно :

$$F_{X,Y}(x, y) = P\{X \leq x, Y \leq y\} \quad -\infty < x, y < \infty$$

Неке карактеристике функције расподеле дводимензионалне случајне променљиве :

1. $0 \leq F_{X_1,X_2}(x_1, x_2) \leq 1$
2. $F_{X_1,X_2}(-\infty, -\infty) = 0$

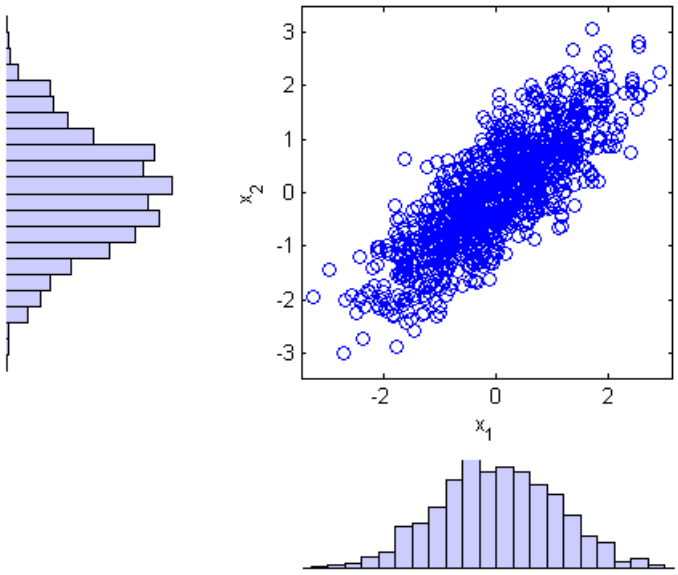
3. $F_{X_1,X_2}(-\infty,-\infty)=0$
   $F_{X_1,X_2}(-\infty,x_2)=0$ $F_{X_1,X_2}(x_1,-\infty)=0$

4. $F_{X_1,X_2}(\infty,\infty)=1$

5.
$$F_{X_1,X_2}(x_1,\infty)=F_{X_1}(x_1) \tag{1.13}$$

6.
$$F_{X_1,X_2}(\infty,x_2)=F_{X_2}(x_2) \tag{1.14}$$

Једнакостима (3.13) и (3.14) су дефинисане **маргиналне расподеле** случајних променљивих X_1 и X_2. Маргиналне расподеле су уствари расподеле једнодиментионалних случајних променљивих X_1 и X_2. На следећој слици је предсављена заједничка расподела две случајне приомењиве заједно са њиховим маргиналним расподелама.

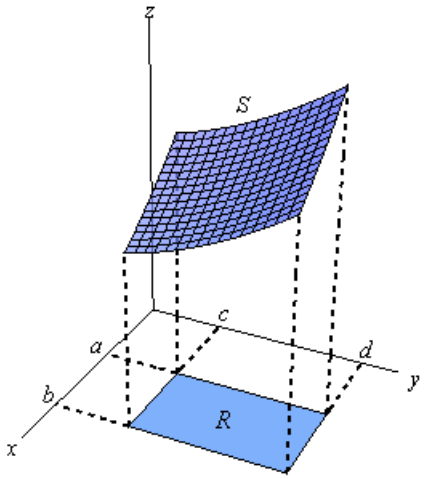


Слика 1.8: Дводимензионална случајна променљива и маргиналне расподеле

Заједничка функција густине (енг. joint density function) дводимензионалне случајне променљиве се дефинише као :

$$f_{X_1,X_2}(x_1,x_2)=\frac{d^2F_{X_1,X_2}(x_1,x_2)}{dx_1dx_2} \tag{1.15}$$

У случају једнодимензионалне случајне променљиве, површина испод графика функције густине на неком интервалу представљала је вероватноћу да случајна променљива узме вредност из тог интервала. У случају дводимензионалне случајне промељиве од интереса је пронаћи вероватноћу да она узме вредност из неке **области**. Та вероватноћа представља **запремину** тела ограниченог функцијом густине са горње стране и датом облашћу са доње стране.



Слика 1.9: Вероватноћа да дводемнзионална случајна промељива (X, Y) узме вредности из области R

Неке особине функције густине :

1. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = 1$

2. $F_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

3. $F_{X_1}(x_1) = \int_{-\infty}^{x_1}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$
   $F_{X_2}(x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

4. $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_2$
   $f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1$

5. $P\{x_{11} < X_1 \leq x_{12}, x_{21} < X_2 \leq x_{22}\} = \int_{x_{21}}^{x_{22}}\int_{x_{11}}^{x_{12}} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

Функција условне расподеле и густине

У неким специфичним случајевима је потребно пронаћи расподелу једне случајне променљиве знајући вредност друге случајне променљиве. Таква расподела назива се **условном расподелом** и обележава се са $F_{X_1}(x_1 \mid X_2 = x_2)$. Аналогно, може се дефинисати и проблем налажења функције густине једне случајне променљиве знајући вредност друге случајне променљиве и таква функција густине се означава са $f_{X_1}(x_1 \mid X_2 = x_2)$. Према [8] условна расподела односно густина се рачуна по следећем обрасцу (детаљно извођење се може наћи у [8])

$$F_{X_1}(x_1 \mid X_2 = x_2) = \frac{\int_{-\infty}^{x_1} f_{X_1,X_2}(x_1,x_2)dx_1}{f_{X_2}(x_2)}$$

односно :

$$f_{X_1}(x_1 \mid X_2 = x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}$$

1.2 Важније расподеле

1.2.1 Бинонма и полиномна(енг. multivariate) расподела

Бинонма расподела

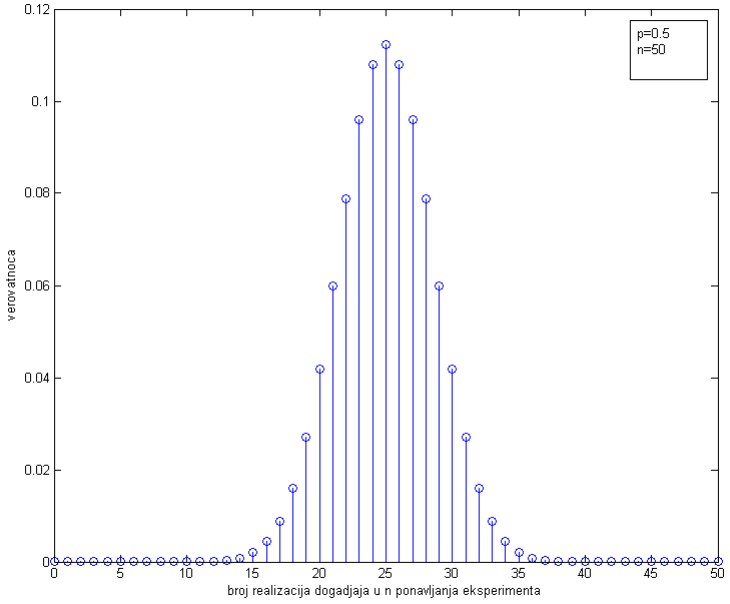
Нека ја A догађај неког екеперимента E који се реализује са вероватноћом $P(A) = p$. Тада је вероватноћа супротног догађаја $P() = 1 - p = q$. Резултат експеримента који је до интереса је остваривање или неостваривање догађаја A. Дакле, може се сматрати да је скуп свих елементарних исхода $\Omega = \{A, \overline{A}\}$. Нека се експеримент понавља **независно** и у неизмењеним условима n пута. На тај начин је формиран **сложени експеримент** чији скуп елементарних исхода садржи све могуће n-торке састављене од и $\overline{A}$ и има их укупно 2^n. Нека је, даље, на том скупу елементарних исхода дефинисана случајна променљива X_n као број остваривања догађаја A у n понављања експеримента E. Вероватоћа да ова случајна променљива узме конкретну вредност k је :

$$p_k = P\{X_n = k\} = \binom{n}{k} p^k q^{n-k}$$

Вероватноће $P\{X_n = k\}, (k = 0, 1, .., n)$ дефинишу **биномну расподелу**, у ознаци $\mathbb{B}(n, p)$. Ова расподела је дискретног типа а њена функција расподеле(кумулативна) се може изразити као :

$$F(x) = \begin{cases} 0 & , x \leq 0 \\ \sum_{k=0}^{r} \binom{n}{k} p^k q^{n-k} & 0 < r < n \\ 1 & x > n \end{cases}$$

Закон расподеле вероватноћа случајне променљиве биномне расподеле приказан је на следећој слици :



Слика 1.10: Биномна расподела - закон расподеле

Полиномна (енг. multivariate) расподела

Изводи се серија од n независних експеримената при чему резултат експеримента може бити један од **коначно много** догађаја : $A_1, A_2, .., A_k, \sum_{i=1}^{k} A_i = \Omega, P(A_i) = p_i (i = 1, 2, ..k)$. Ако се дефинише k-диментионална случајна применљива $(S_n^{(1)}, ..., S_n^{(k)})$, где $S_n^{(i)}$ представља број релаизација случајног догађаја A_i у n независних експеримената, тада важи :

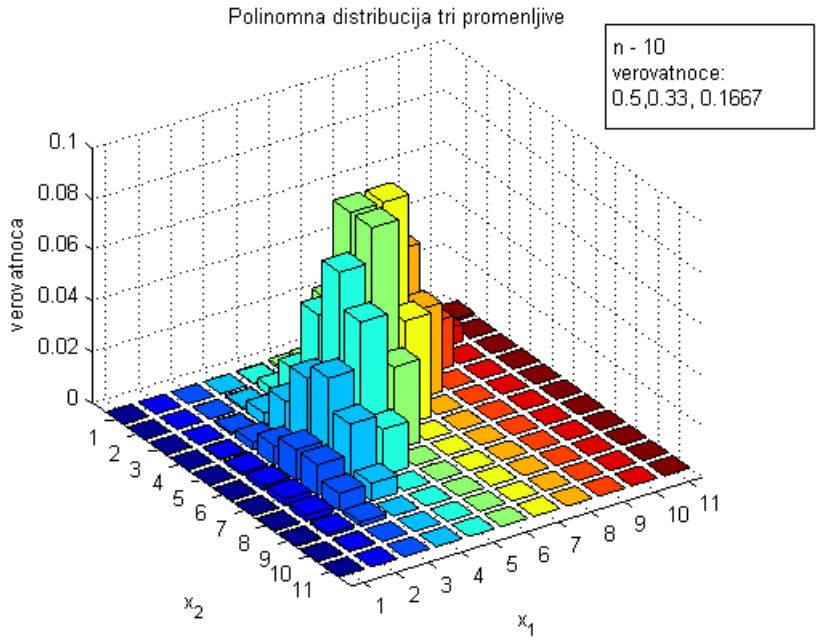
$$P(S_n^{(1)} = r_1, .., S_n^{(k)} = r_k) = \frac{n!}{r_1!..r_k!} p_1^{r_1} ... p_k^{r_k}$$

$$r_1, ..r_k \in \{0, 1, .., n\} \quad r_1 + .. + r_k = n$$

Ако се са $S = (S_n^{(1)}, ..., S_n^{(k)})$ означи k-диментионална случајна применљива која има полиномијалну расподелу тада се то записује као :

$$S \sim Mult(n, p)$$

где је $p = (p_1, p_2, .., p_k)$ Пример полиномне расподеле при чему резулатат експеримента може бити један од **три** догађаја, дат је на следећој слици :



Слика 1.11: Биномна расподела - закон расподеле

1.2.2 Дирихлеова расподела

Дирихлеова расподела представља фамилију расподела **за параметре** p полиномијалне расподеле. Задаје се са :

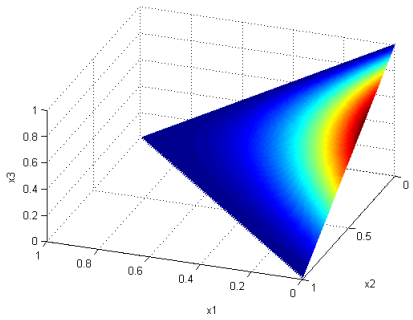
$$Dir(p; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

при чему је α параметар расподеле а B означава мултиномијалну *бета функцију*. Мултиномијалну бета функција се изражава преко гама фунцкије на следећи начин

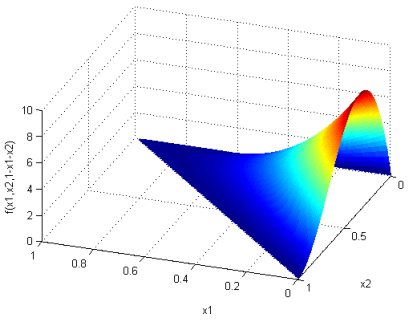
$$B(\alpha) = \frac{\prod_{i=1}^{|\alpha|} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{|\alpha|} \alpha_i)}$$

На слећој слици је графички представљена Дирихлеова расподела за три променљиве :

- $\alpha = (1, 2, 3)$

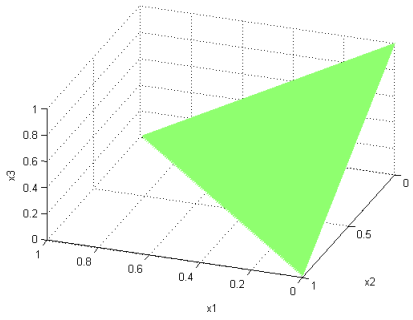


Слика 1.12: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу

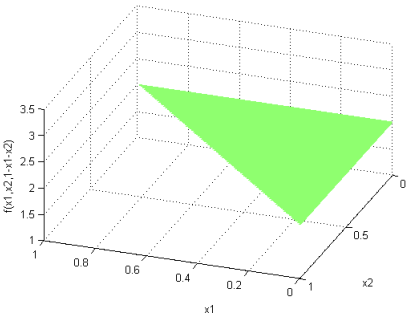


Слика 1.13: Дирихлеова расподела у три димензије

- $\alpha = (1, 1, 1)$

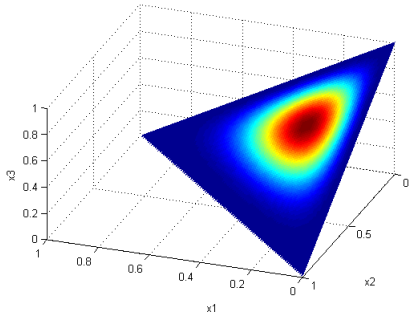


Слика 1.14: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу

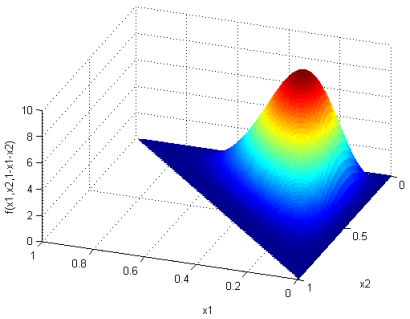


Слика 1.15: Дирихлеова расподела у три димензије

- $\alpha = (3, 3, 5)$



Слика 1.16: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу



Слика 1.17: Дирихлеова расподела у три димензије

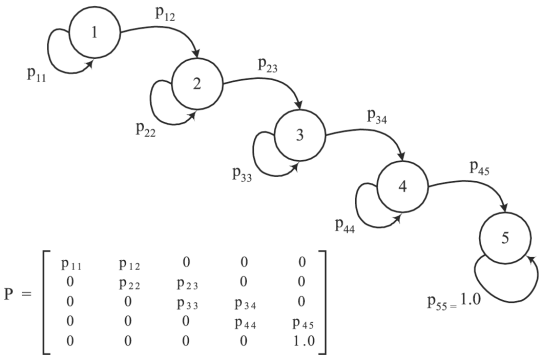
1.3 Гибсово узорковање

1.3.1 Марковљеви ланци

Марковљевим ланцима моделује се математички сиситем стања и прелаза међу тим стањима.

Дефиниција 9 *Случајан (стохастички) процес представља математички модел процеса чија је еволуција описана законима вероватноће.*
Марковљеви процеси су они случајни процеси чије будуће стање зависи само од тренутног стања. Оваква особина још се назива и одсуство памћења
Марковљеви ланци представљају посебну врсту Марковљевих процеса где се процес може налазити само у коначном броју стања.

Пример Марковљевог ланца дат је на следећој слици :



Слика 1.18: Марковљев ланац - графички пример и матрица транзиције

Систем се састоји од 5 стања. У сваком стању, са одређеним вероватноћама систем може да пређе у неко од следећих стања - конкретно да остане у тренутном стању или да пређе у једно стање ниже. Вероватноћа преласка у следеће стање зависи само од тренутног стања. Марковљеви ланци се често представљају **матрицама транзиције** при чему i-та врста у матрици садржи вероватноће преласка у свако од стање система када се систем налази у стању i. Сума свих вероватноћа у свакој врсти је 1 (систем сигурно мора да се нађе у неком стању, дакле вероватноћа да систем пређе у неко стање, могуће исто, је 1). Свака врста представља условни закон расподеле вероватноћа да систем пређе у било које стање у односу на тренутно стање (i-та врста - i-то стање). Сака колона представља маргиналну расподелу вероватноћа да се систем нађе у одређеном стању (i-та колона - i-то стање).

Систем се у једном тренутку може налазити у само једном стању. Нека се стање система карактерише случајном премнљивом X_n која у тренутку n има расподелу $\overrightarrow{s}$ и нека је укупан број стања система M. Расподела $\overrightarrow{s}$ је у ствари закон расподеле (енг. PMF) јер се ради о случајној променљивој дискретног типа - систем може бити само у једном од M могућих стања, и конкретно може се посматрати као вектор врсте, димензија $1 \times M$ где се на i - том месту налази вероватноћа да се систем у тренутку n нађе у стању i.

У следећем временском тренутку, $n+1$, систем се може наћи у било ком од M стања са различитим вероватноћама. Вероватноћа да ће се систем у тренутку $n+1$ наћи у стању j означава се са $P(X_{n+1} = j)$. Пошто ова вероватноћа зависи од стања у претходном тренутку, може се изразити на следећи начин (према формули тоталне вероватноће)

$$P(X_{n+1} = j) = \sum_{i}^{M} P(X_{n+1} = j \mid Xn = i)P(Xn = i) = *$$

$P(X_{n+1} = j \mid Xn = i)$ = вероватноћа преласка система из стања i у стање j -> $p_{i,j}$
$P(Xn = i)$ = вероватноћа да се систем у тренутку n нађе у стању i -> s_i

$$* = \sum_{i}^{M} p_{i,j} s_i$$

Дакле, верованоћа да систем у $n+1$-ом тренутку буде у стању j једнака је суми производа вероватноћа да се систем у n-том тренутку нађе у било ком стању и вероватноћа одговарајућих прелаза.

Ова сума представља j-ту колону у матрици (димензија $1 \times M$) која се добије при множењу вектора $\overrightarrow{s}$ и матрице транзиције P.

Према свему наведеном следи да је закон расподеле случајне променљиве X_{n+1} (расподела вероватноћа да се систем у $n+1$-ом тренутку налази у сваком од стања) једнак $\overrightarrow{s} \times P$.

Аналогно, у тренутку n+2, случајна променљива X_{n+2} има расподелу $\overrightarrow{s} \times P^2$, у тренутку $n+3$, случајна променљива X_{n+3} има расподелу $\overrightarrow{s} \times P^3$ итд.

Дефиниција 10 *Расподела $\overrightarrow{s}$ за коју важи :*

$$\overrightarrow{s} \times P = \overrightarrow{s}$$

*назива се **стационарна** или **равнотежна** расподела.*

$\overrightarrow{s} \times P$ представља "један корак у будућност", тј. расподелу вероватноћа да систем нађе у сваком од стања у следећем временском тренутку. Уколико расподела остаје иста, односно, вероватноће се не мењају са временом, тада се та расподела назива стационарном. Под одређеним условима везаним Марковљеве ланце, доказује се да Марковљев ланац **увек** конвергира ка својој стационарној расподели без обзира на полазно стање. Више о конвергенцији Марковљевих ланаца може се наћи у [11]. Дакле, полазне стање се може изабрати потпуно случајно а затим, уколико се дозвволи да "протекне"довољно времена, закон расподеле вероватноће да се систем нађе у свим стањима система ће конвергирати ка стационарној расподели тог ланца.

Дефиниција 11 *MCMC (енг. Markov Chain Monte Carlo) методе представљају класу алгоритама који се користе за синтетичко генерисање узорака случајних променљивих из одговрајућих расподела. Овим методама се креирају Мерковљеви ланаци који као равнотежну расподелу имају расподелу из које се узимају узорци. Једна од MCMC метода је и Гибсово узорковање (енг. Gibbs sampling)*

Гибсово узорковање

Нека је дата заједничка расподела (енг. joint distribution) $p(\mathbf{z}) = p(z_1, z_2, .., z_M)$ из које је потрбено одабрати неку вредност (енг. sample) и нека је познато почетно стање Марковљевог ланца који је потребно генерисати. Сваки корак Гибсовог узорковања почиње заменом вредности једне променљиве $z_1, z_2, .., z_M$ вредношћу која се добија из **условне расподеле** те променљиве у односу на све остале. Дакле, z_i се мења вредношћу која се узима из расподеле $p(z_i \mid z_{-i})$, где је са z_i означена i-та координата вектора z а са z_{-i} сви $z_1, z_2, .., z_M$ без z_i. Ова процедура се наставља за све променљиве по неком одређеном редоследу. При довољном броју итерација, врсности вектора z ће конвергирати ка $p(z)$.

На пример, нека је дата расподела три случајне променљиве $p(z_1, z_2, z_3)$ и нека су вредности у тренутку t : z_1^t, z_2^t, z_3^t. Нека се замена вредности променљивих врши у односу на индекс, од најмањег ка највећем. Вредност z_1^t се мења новом вредношћу z_1^{t+1} која се узима (узрокује) из расподеле

$$p(z_1 | z_2^t, z_3^t).$$

Сада се вредност z_2^t мења са вредношћу z_2^{t+1} која се узима из расподеле

$$p(z_2|z_1^{t+1}, z_3^t).$$

Дакле, одмах се користи нова вредност промењиве z_1. Коначно, за промену вредности z_2^t користи се вредност z_3^{t+1} која се добија из расподеле :

$$p(z32|z_1^{t+1}, z_2^{t+1}).$$

Овим је завршена **једна итерација** Гинбсовог узорковања. Исти процес се наставља кроз низ итерација све до одређеног броја или до неког другог услова заустављања.

Описана процедура се може уопштити и на више од три променљиве и може се представити следећим псеудокодом:

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

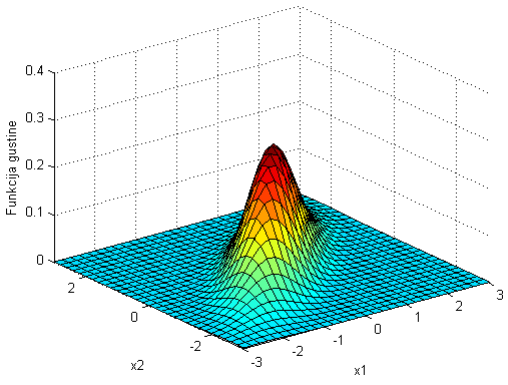
Слика 1.19: Псеудокод Гибсовог узорковања, преузето са [12]

Гибсово узорковање подразумева да су унапред познате **условне расподеле** свих променљивих и да је могуће узорковање из њих.

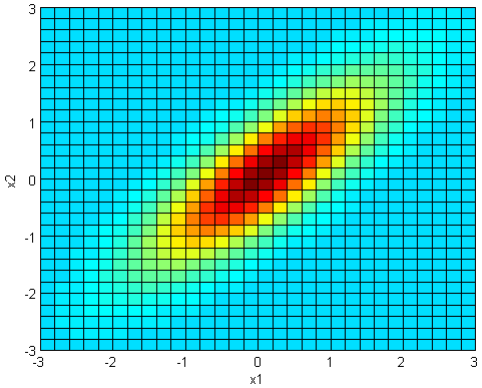
Пример : Нека је потребно узорковати вредности из дводимензионалне нормалне расподеле $\mathcal{N}(\mu, \Sigma)$ Гибсовим узорковањем при чему је

$$\mu = [\mu_1, \mu_2] = [0, 0]$$
$$\Sigma = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Графички приказ овакве дводимензионалне нормалне расподеле дат је на следећој слици (тродимензионално и пројектовано на две димензије):



Слика 1.20: Тродимензионални приказ дводимензионалне нормалне расподеле



Слика 1.21: Дводимензионлни приказ дводимензионалне нормалне расподеле

Основна претпоставка Гибсовог узорковања је да су познате условне расподеле свих променљивих и да је из њих могуће узорковати. Према [12] и [13], за условне расподеле дводимензионалне заједничке расподел важи :

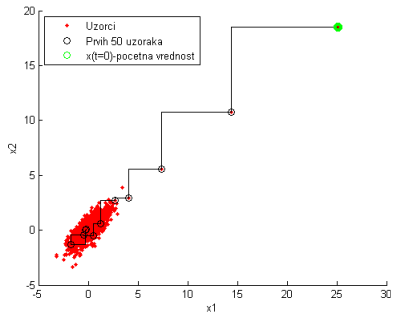
$$p(x_1 \mid x_2^{(t-1)}) = \mathcal{N}(\mu_1 + \rho_{21}(x_2^{(t-1)} - \mu_2), \sqrt{1 - \rho_{21}^2})$$

и

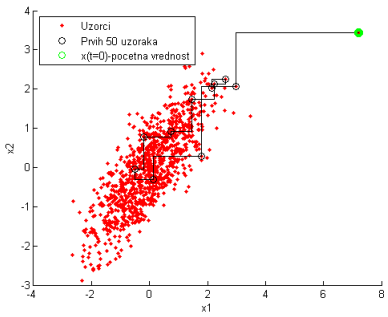
$$p(x_2 \mid x_2^{(t)}) = \mathcal{N}(\mu_2 + \rho_{12}(x_2^{(t)} - \mu_2), \sqrt{1 - \rho_{12}^2})$$

Дакле, обе условне расподеле представљају једнодимензионалну нормалну расподелу са одговарајућим параметрима. Почетне вредности променљивих се бирају случајно зато што нису од важности. Марковљев ланац ће свакако конвергирати ка дводимензионалној нормалној расподели са неведеним параметрима после одређеног броја итерација. У зависности од полазног стања, тај број итерација ће бити мањи или већи.

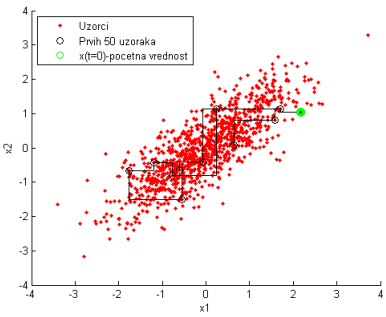
На следећем сликама су предстваљене добијене расподеле Гибсовим узорковањем за различите почетне вредности :



Слика 1.22: Почетна тачка (25.1284,18.5165)



Слика 1.23: Почетна тачка (7.2162, 3.4380)



Слика 1.24: Почетна тачка (2.1649, 1.0314)

Са слика је очигледно да се првих неколико узорака може занемарити (3.22 може се занемарити првих 7-8 узорака).

Важно је приметити да се узимање узорака увек креће по "степенастом"обрасцу. Дакле, две суседне тачке имају исту једну координату (x или y). То је зато што Гибсово узорковање у једном тренутку мења **само једну** променљиву у односу на одговарајућу вредност друге.

Конвергенција алгоритма Гибсовог узорковања ка стационарној расподели Марковљевог ланца је теоретски загарантована, али је у пракси јако тешко одредити број итерација након којих ланац почиње да конвергира. Један од начина процене конвергенције је и рачунање *log-likelihood* -а

1.4 Како ради TM алгоритам

Раније је неформално описан LDA генеративни процес. Основна претпоставка је да се сваки документ у одређеној пропорцији говори о свакој теми (има одређену расподелу над темама) као и да свакој теми све речи из корпуса припадају са различитим вероватноћама (расподела над речима). Генеративни процес се , према [14], може описати следећим псеудокодом :

1. For $k = 1...K$:
   (a) $\phi^{(k)} \sim Dirichlet(\beta)$
2. For each document $d \in \mathbf{D}$:
   (a) $\theta_d \sim Dirichlet(\alpha)$
   (b) For each word $w_i \in d$:
      i. $z_i \sim Discrete(\theta_d)$
      ii. $w_i \sim Disctete(\phi^{(z_i)})$

Слика 1.25: Генеративни процес LDA-а

при чему је : K - укупан број тема у колекцији $\phi_{(k)}$ - расподела над свим речима из колекције и представља расподелу над речима у k-тој теми θ_d - расподела над темама у документу d. z_i - тема којој припада реч w_i. α, β - **хиперпараметри** тј. параметри симетричних Дирехлеових расподела.

Описани генеративни процес резултује формирањем следеће заједничке расподеле :

$$p(w, z, \theta, \phi \mid \alpha, \beta) = p(\phi \mid \beta)p(\theta \mid \alpha)p(z \mid \theta)p(w \mid \phi_z) \tag{1.16}$$

Непозанте променљиве које је потребно "открити"су z, θ и ϕ на основу (једино) познатих **речи** и њиховог присуства у сваком од докумената. Дакле, потребно је пронаћи расподеле наведених променљивих **под условом** да су познате речи и њихова распоређеност по документима тј. открити њихове постериорне расподеле. Основни проблем ТМ је **постериорно закључивање** (енг. posterior inference) односно отривање постериориних расподела латентних случајних променљивих на основу задатог скупа докумената и речи што представља решавање следеће једначине :

$$p(\theta, \phi, z \mid w, \alpha, \beta) = \frac{p(\theta, \phi, z, w, \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \tag{1.17}$$

Према [14], рачунање имениоца овог разломка је готово немогуће па се стога прибегава апроксимативним методама каква је и Гибсово узорковање.

Да би се применило Гибсово узорковање, потребно је познавати условне расподеле свих променљивих из чије се заједничке расподеле узоркује. Међутим, показује се да је довољно пронаћи само z јер се остале две променљиве могу преко ње израчунати и то (према [14]):

$$\theta_{d,z} = \frac{n(d, z) + \alpha}{\sum_{|Z|} n(d, z) + \alpha}$$

$$\phi_{z,w} = \frac{n(z, w) + \beta}{\sum_{|W|} n(z, w) + \beta}$$

Овако примењен алгоритам Гибсовог узорковања назива се још и енг. Collapsed Gibbs Sampling. Дакле, циљ је пронаћи за сваку реч, вероватноћу да припадне свакој од тема, под условом да су познате теме којима припадају остале речи у том тренутку. Формалније, ово се може записати $p(z_i \mid z_{-i}, \alpha, \beta, w)$ где z_{-i} представља доделу тема свим речима сем i-те.

Априорне расподеле коришћене у ТМ су Дирихлеове.Важна особина Дирихлеове расподеле је да је она **конјугована** са мултиномијалном расподелом. Дирихлеова расподела је расподела над параметрима мултиномијалне расподеле. Нека је на почетку претпостављено да параметри мултиномијалне расподеле припадају некој Дирихелеовој расподели - $\mathbf{p} \sim Dir(\mathbf{p}, \alpha)$. Нека је $\mathbf{x}$ узорак генерисан из мултиномијалне расподеле $Mult(\mathbf{x}; \mathbf{p})$. Тада важи да је постериорна

расподела $\mathbf{p}$ (дакле, расподела под условом да је познат узорак $\mathbf{x}$) такође **Дирихлеова расподела** са параметром $\mathbf{x}+\alpha$ тј.:

$$p(\mathbf{p} \mid \mathbf{x}, \alpha) = Dir(\mathbf{p}; \mathbf{x}+\alpha) = \frac{1}{B(\mathbf{x}+\alpha)} \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} \tag{1.18}$$

Како (3.18) представља **расподелу** то важи да је :

$$1 = \int \frac{1}{B(\mathbf{x}+\alpha)} \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} = \frac{1}{B(\mathbf{x}+\alpha)} \int \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} \tag{1.19}$$

Одакле следи да је :

$$\int \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} = B(\mathbf{x}+\alpha) \tag{1.20}$$

Ова једнакост је важна за даљи опис рада TM алгоритма.

Према формули условне расподеле, важи :

$$p(z_i \mid z_{-i}, \alpha, \beta, w) = \frac{p(z_i, z_{-i}, w \mid \alpha, \beta)}{z_{-i}, w \mid \alpha, \beta)} \propto p(z_i, z_{-i}, w \mid \alpha, \beta) = p(z, w \mid \alpha, \beta) \tag{1.21}$$

$p(z, w \mid \alpha, \beta)$ се може посматрати као "маргиналан расподела"две променљиве заједничке расподеле (3.16) па важи :

$$p(z, w \mid \alpha, \beta) = \iint p(z, w, \theta, \phi \mid \alpha, \beta) d\theta d\phi = \iint p(\phi \mid \beta) p(\theta \mid \alpha) p(z \mid \theta) p(w \mid \phi_z) d\theta d\phi \tag{1.22}$$

Груписањем по заједночкој зависној променљивој, претхонда једначина се може написати :

$$p(z, w \mid \alpha, \beta) = \int p(z \mid \theta) p(\theta \mid \alpha) d\theta \int p(w \mid \phi_z) p(\phi \mid \beta) d\phi \tag{1.23}$$

Оба интерграла представљају комбинацију узорка из мултиномијалне расподеле и априорне Дирихлеове расподеле. Како је Дирихлеова расподела конјугована (conjugate prior) са мултиномијалном, у подинтегралном изразу се налази "множење"две Дирихлеове расподеле са одговарјућим параметрима.

Дакле : Пошто $p(z \mid \theta)$ има мултиномијалну дистрибуцију, важи:

$$p(z \mid \theta) = \prod_{i=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{\Omega_{d,k}} \tag{1.24}$$

, где је $\Omega_{d,k}$ број који означава колико пута је тема k додељена у документу d - број речи који у документу d припадају теми k.

Члан $p(\theta \mid \alpha)$ је из основне Дирихлеове расподеле па важи :

$$p(\theta \mid \alpha) \overset{(1)}{=} \prod_{i=1}^{D} p(\overline{\mathbf{q_d}} \mid \alpha) \overset{(2)}{=} \prod_{d-1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} q_{d,k}^{\alpha_k-1} \tag{1.25}$$

где је $\overline{\mathbf{q_d}}$ расподела вероватноћа тема у документу d. Расподеле вероватноћа тема по документима су независне, па је зато могуће написати (1).Расподела тема по документу се узима из Дирихлеове расподеле па је зато могуће написати (2).

Према томе, први интеграл једнакости (3.23) се записује као :

$$\int p(z \mid \theta)p(\theta \mid \alpha)d\theta = \int \prod_{i=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{\Omega_{d,k}}\prod_{d-1}^{D}\frac{1}{B(\alpha)}\prod_{k=1}^{K}q_{d,k}^{\alpha_k-1}d\theta_d \overset{(1)}{=} \prod_{i=1}^{D}\int \frac{1}{B(\alpha)}\prod_{k=1}^{K}q_{d,k}^{\Omega_{d,k}+\alpha_k-1}d\theta_d \tag{1.26}$$

Једнакост (1) следи из чињенице да су θ_d независне расподеле па се могу интегралити посебно - правило интеграције производа Према релацији (3.20) претходна једнакост се може написати и као се :

$$\int p(z \mid \theta)p(\theta \mid \alpha)d\theta = \prod_{i=1}^{D}\frac{B(\Omega_d+\alpha)}{B(\alpha)} \tag{1.27}$$

где је са Ω означена матрица докумената и тема, $\Omega_{d,k}$ означава колико је пута тема k додељена у документу d а Ω_d је d-та врста те матрице. Елементи ове матрице могу се математички записати и овако :

$$\Omega_{d,k} = \sum_{i=1}^{N} I(d_i = d \wedge z_i = k) \tag{1.28}$$

где је N укупан број речи у корпусу(са понављањем).

Аналогно претходним извођењима, и други интеграл може да се упрости:

Члан $p(\phi \mid \beta)$ је из основе Дирхлеове расподеле па важи :

$$p(\phi \mid \beta) = \prod_{k=1}^{K} p(\phi_k \mid \beta) = \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\beta_v-1} \tag{1.29}$$

Члан $p(w \mid \phi_z)$ има мултиномијалну расподелу па важи :

$$p(w \mid \phi_z) = \prod_{i=1}^{N} p(\phi_{z_i,w_i}) = \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{\Psi_{k,v}} \tag{1.30}$$

где је $\psi \quad K \times V$ матрица а $\psi_{k,v}$ броји колико тема k била додељена речи v. Ова матрице може се још написати као :

$$\psi_{k,v} = \sum_{i=1}^{N} I(w_i = v \wedge z_i = k) \tag{1.31}$$

На основу (3.29) и (3.30) следи :

$$\int p(w \mid \phi_z)p(\phi \mid \beta)d\phi = \int \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\psi_{k,v}+\beta_v-1}d\phi_k \tag{1.32}$$

Аналогно извођењу (3.26) (3.27) следи :

$$\int \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\psi_{k,v}+\beta_v-1}d\phi_k = \prod_{k=1}^{K}(\frac{1}{B(\beta)}\int \prod_{v=1}^{V}\phi_{k,v}^{\psi_{k,v}+\beta_v-1}d\phi_k) = \prod_{k=1}^{K}\frac{B(\psi_k+\beta)}{B(\beta)} \tag{1.33}$$

Коришћењем једнакости (3.27) и (3.33), једнакост (3.23) се може записати као :

$$p(z, w \mid \alpha, \beta) = \prod_{i=1}^{D}\frac{B(\Omega_d+\alpha)}{B(\alpha)}\prod_{k=1}^{K}\frac{B(\psi_k+\beta)}{B(\beta)} \tag{1.34}$$

На основу (3.34) може се извести правило по коме ће Гибсов алгоритам узорковања мењати доделе тема речима. Дакле :

$$p(z_i = k|Z^{-i}, W, \alpha, \beta) = \frac{p(z_i = k, Z^{-i}, W \mid \alpha, \beta}{p(Z^{-i}, W \mid \alpha, \beta)} = \frac{p(Z, W \mid \alpha, \beta)}{p(Z^{-i}, W \mid \alpha, \beta)} \tag{1.35}$$

Именилац претходне једноакост се може написати преко условне вероватноће у следећем облику :

$$p(Z^{-i}, W \mid \alpha, \beta) = p(Z^{-i} \mid \alpha\beta)p(W \mid Z^{-i}, \alpha\beta) \overset{(1)}{=} \tag{1.36}$$

$$= p(Z^{-i})p(W^{-i} \mid Z^{-i})p(w_i) \propto p(Z^{-i})p(W^{-i} \mid Z^{-i}) = p(Z^{-i}, W^{-i}) \tag{1.37}$$

Једнакост (1) следи из чињенице да свако z_i зависи само од w_i. Од ове једнакости параметри α, β су изостављени због прегледности, али се подразумевају.

Форма једнакости (3.37) иста је као (3.31) па се једнакост (3.35) записује као :

$$p(z_i = k|Z^{-i}, W, \alpha, \beta) = \prod_{k=1}^{K} \frac{B(\psi_k + \beta}{B(\psi_k^{-i} + \beta} \prod_{d=1}^{D} \frac{B(\Omega_d + \alpha}{B(\Omega_d^{-i} + \alpha} \tag{1.38}$$

Коришћењем особина бета фунцкије, претходна једнакост се своди на :

$$p(z_i = k|Z^{-i}, w = v, W^{-i}, \alpha, \beta) = \frac{\psi_{k,v} + \beta_{w_i} - 1}{\left[\sum_{v=1}^{V} \psi_{k,v} + \beta_v\right] - 1} \left[\Omega_{d,k} + \alpha_k - 1\right] \tag{1.39}$$

Детаљно извођење може се наћи код [14] и [15].

1.4.1 Имплементација - псеудокод

Овде ће доћи мој псеудокод, ово је само привремено

Input: words $\mathbf{w} \in$ documents $\mathbf{d}$
Output: topic assignments $\mathbf{z}$ and counts $n_{d,k}, n_{k,w}$, and n_k
begin
  randomly initialize $\mathbf{z}$ and increment counters
  foreach *iteration* **do**
    for $i = 0 \to N - 1$ **do**
      $word \leftarrow w[i]$
      $topic \leftarrow z[i]$
      $n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1
      for $k = 0 \to K - 1$ **do**
        $p(z = k|\cdot) = (n_{d,k} + \alpha_k)\frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$
      end
      $topic \leftarrow$ sample from $p(z|\cdot)$
      $z[i] \leftarrow topic$
      $n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1
    end
  end
  return $\mathbf{z}$, $n_{d,k}, n_{k,w}, n_k$
end

Algorithm 1: LDA Gibbs Sampling

Слика 1.26: Псеудокод - преузето са [15]

Литература

[1] Lars Vogel, Android Service and Broadcast Receiver, www.vogella.de, 2011.

[2] David M. Blei,Introduction to Probabilistic Topic Models,Princeton University

[3] David M. Blei, Topic Models,Princeton University,September 1,2009

[4] David Mimno,The details: training and validating big models on big data,Princeton University

[5] Ivana Kovačević, Verovatnoća i statistika sa zbirkom zadataka, Beograd 2011.

[6] Violeta Aleksić, Elementi teorije verovatnoće i matematičke statistike,

[7] www.ekfak.kg.ac.rs, kurs Osnovi statistike, avgust 2015.

[8] http://starisajt.elfak.ni.ac.rs/phptest/new/html/Studije/predavanja-literatura/matematika-odabrana-poglavlja/verovatnoca.pdf, avgust 2015

[9] Random Signals and Processes Primer with MATLAB, Gordana Jovanovic Dolocek,2013

[10] Lecture 32: Markov Chains Continued | Statistics 110 on youtube - https://www.youtube.com/watch?v=aBGOyZv2pZE, avgust 2015

[11] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

[12] http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html, avgust 2015

[13] https://theclevermachine.wordpress.com/2012/11/05/mcmc-the-gibbs-sampler/, avgust 2015

[14] William M. Darling, A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling, December 1, 2011

[15] Yi Wang, Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details, August , 2008